Semantic Parsing with Verbal Subcategorization

Jordi Atserias
Irene Castellón
Montse Civit
German Rigau

The aim of this work is to explore new methodologies on Semantic Parsing for unrestricted texts. Our approach follows the current trends in Information Extraction (IE) and is based on the application of a verbal subcategorization lexicon (LEXPIR) by means of complex pattern recognition techniques. LEXPIR is framed on the theoretical model of the verbal subcategorization developed in the Pirapides project.

1 Introduction

Most of the different tasks included in Natural Language Processing (such as Information Retrieval, Information Extraction, Information Filtering, Natural Language Interfaces and Story Understanding) apply different levels of Natural Language Understanding. For instance, in the case of Information Extraction, the Natural Language Understanding component plays a crucial role. This is due to the fact that most of the information to be extracted can only be identified by recognizing the conceptual roles. This area has been greatly promoted by the Message Understanding Conferences (MUC's) organized by TIPSTER.

Such conferences have shown the tendency of the Information Extraction Systems to be more domain [Wilks and Catizone, 1999] and language independent [Humphreys et al., 1998] [Kilgariff, 1997], making Information Extraction stand closer to Natural Language Understanding. Currently, other related areas (such as Story Understanding [Riloff, 1999]) have begun to adapt the recent improvements done in Information Extraction.

An important step in any process that implies Natural Language Understanding is Semantic Interpretation. Semantic Interpretation can be defined as the process of obtaining a suitable meaning representation for a text. The input of the Semantic Interpreter can vary largely, going from raw text to full parsing trees. Likewise, the output of the Semantic Interpreter can also vary considerably (logical formulae, case-frames, SQL), mostly influenced by the type of application. In relation to this, two important sub-tasks can be distinguished within Semantic Interpretation: Word Sense Disambiguation (WSD) and Semantic Parsing, being the latter the interest of the current work. Further, an essential part of the Semantic Parsing involves the production of a case-role analysis in which the semantic roles of the entities, such as *starter* or *instrument*, are identified [Brill and Mooney, 1997].

The work here presented focuses on this problem, in particular on the issue of obtaining the verbal argument structure of the sentence. Our proposal for obtaining the representation of the meaning components (roles) of the verb is based on the application of the linguistic theory of the verbal subcategorization developed inside the Pirapides Project [Fernández et al., 1999], and is performed by means of complex pattern recognition techniques.

Pirapides is a project centered on the study of the English, Spanish and Catalan verbal predicates. Pirapides has several goals: On the one hand, from a theoretical point of view, a deep study is being carried out of the units that the model of a verbal entry has produced. This syntactic component focuses on the representation of the interaction between the syntactic and semantic components.

On the other hand, from an application-oriented point of view, a lexicon (LEXPIR) is being developed, based on this theoretical model, which will be used to analyze the corpus.

Following this brief introduction, Section 2 presents the linguistic model and Section 3 the computational model. Then, Section 4 describes the experiments carried out and the results obtained. Finally, Section 5 draws some conclusions and presents further work.

2 Lexical Model

The syntactic analysis using Context Free Grammars (CFG) for non domain-specific Spanish corpora has several limitations: it is basically impossible to carry out an analysis at a sentence level including syntactic functions. This is mainly due to the optionality of some constituents (such as the subject), and also because of the free order of the constituents.

Further, phrase analysis is not enough in order to obtain a suitable interpretation of the sentence. Thus, it becomes necessary to explore new tools to go beyond the phrase level.

Bearing this goal in mind, a hierarchical verbal lexicon for Spanish (LEXPIR) is being developed. In this lexicon, verbs are grouped hierarchaly based on their meaning components as well as their diathetic alternations [Fernández and Martí, 1996] [Fernández et al., 1999], [Morante et al., 1998]. Moreover, each group is subclassified according to the number of components which can be explicitly realized. In addition, LEXPIR includes, for each verb sense, information about the number of arguments, their syntactic realization, the prepositions they can take, their semantic component, their agreement and their optionality.

The information is propagated within the hierarchy in a top-down manner, that is, each verb inherits the elements from its group and each group from its class. However, the inherited information can be overwritten by the information already associated to the specific verb entry (default monotonic inheritance).

N^o Id	**Syntax**	**Prep.**	**Component**	**Semantics**	**Agree.**	**Opt.**
1	NP	p_inic	starter	Human	yes	yes
2	x	x	entity	Top	no	yes
3	PP	p_rut	route	Top	no	yes
4	PP	p_orig	source	Top	no	yes
5	PP	x	destination	Top	no	yes

Table 1: Basic Model for trajectory verbs

As shown in Figure 1, the trajectory class has five components: *starter*(1), *entity*(2) and the *trajectory*, which is the component which defines the class. The trajectory can be further divided into three components: *route*(3), *source*(4) and *destination*(5). Each one of these components has a basic phrase structure, a set of prepositions introducing them and a particular semantics. Moreover, one of the components must be in agreement with the verb.

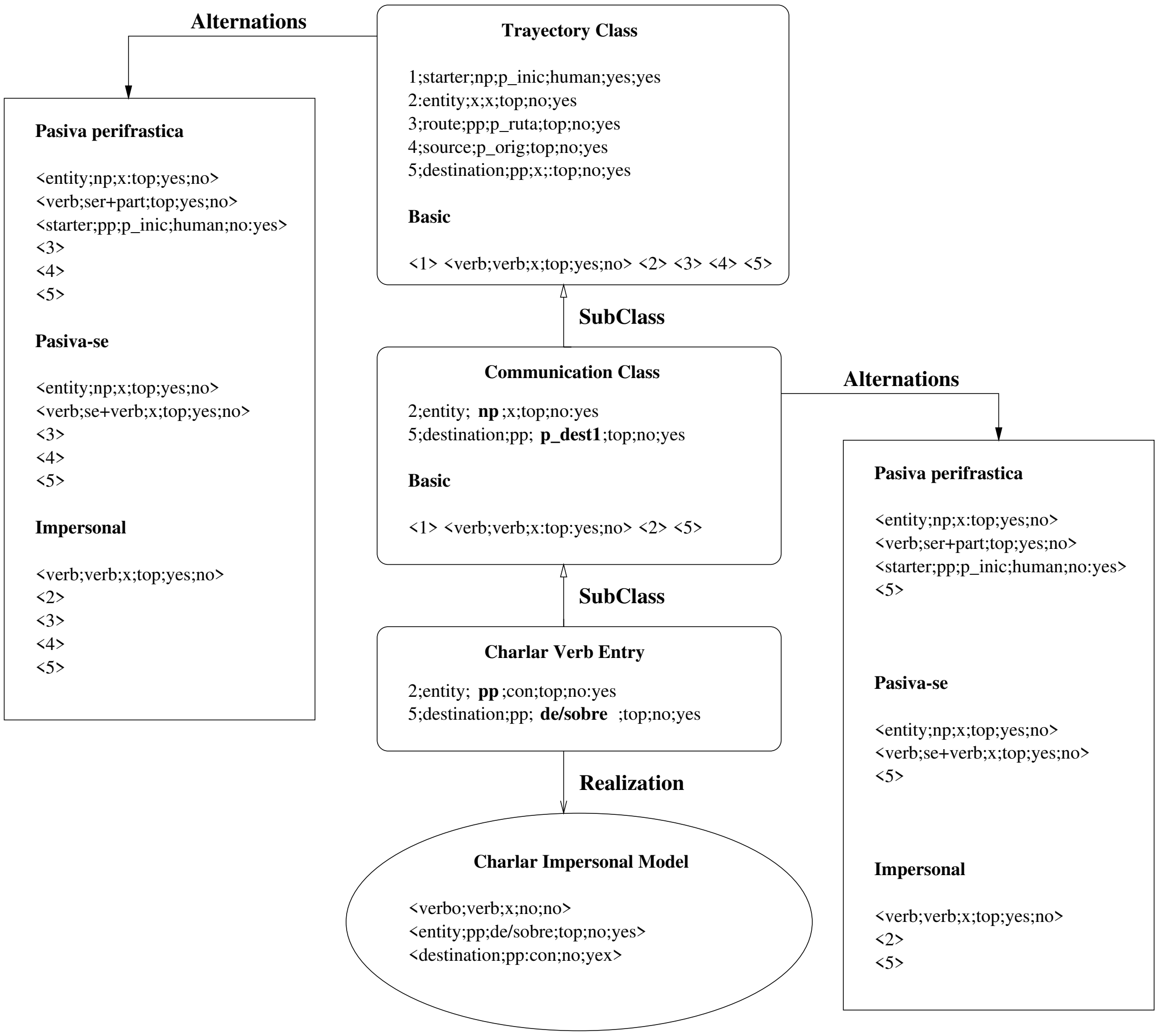

Figure 1: Class Hierarchy

There are four subclasses included in the trajectory class: *non-autonomous movement*, *autonomous movement*, *communication* and *transfer* (although the last one has not been formalized yet).

The non-autonomous movement subclass is characterised by the fact that it explicitly realizes the five components: “Alguien (1) desplaza algo (2) por un lugar (3) desde un punto (4) a otro (5)” (somebody moves something through a place from one point to another).

The autonomous movement presents a co-indexation between the components *starter* and *entity*. Eg.: "Alguien (1,2) va por un lugar (3) desde un punto (4) a otro (5)" (somebody goes through a place from one point to another). As it can be seen in the example, (1) is at the same time the *starter* of the event *go* and the *entity* that is moved.

Finally, in the communication class there are only three components which are explicit: *starter*, which is at the same time the *source*, *entity* and *destination*). Eg.: "Alguien (1,4) dice algo (2) a alguien (5)" (somebody says something to somebody).

Regarding prepositions, those which can appear in the destination component (5) are specified in the subclasses and can be divided into two groups: "p_dest1" which includes "a/para" (to) and "p_dest2" which includes the rest of preposition for *destination*.

Moreover, specific verb forms can impose their own restrictions. For instance, "charlar" (to chat) is a verb which, in contradiction with the rest of the verbs in the communication class, does not accept an NP in the entity component and the PP must have the preposition "de/sobre" (about). Furthermore, it cannot take the prepositions "a/para" to express destination and uses the preposition "con" instead. Eg.: "Alguien (1) charla de algo (2) con alguien (5)" (somebody chats about something with somebody).

Finally, in order to obtain the alternation schemes for a verb, the information of the verb is composed with the alternations of the class. The different elements that appear on a model are explained below for a specific case: the basic model for the trajectory verbs (see Table 1).

- *Id Number*: Numeric value that identifies the meaning component.
- *Syntax*: Syntactic realization of the semantic component. For the second component this information is unspecified (x) as the syntactic realization depends on the subclass. Moreover, this element, which is usually the Direct Object, has other restrictions: if its semantics indicates that it is [+human/animate] it should be a PP, while if it is [-human/animate] it has to be realized as an NP.
- *Preposition*: List of prepositions which have been established according to their meanings and occurrences.
- *Component*: Meaning component determined by the class.
- *Semantics*: Semantics of the component; this is a feature specific of the argument.

- *Agreement*: Person and number agreement with the verb.
- *Optionality*: This indicates which elements are optional inside the sentence.

Treating the optionality of the meaning components within the model itself allows us to reduce the number of possible alternations which have been established at a theoretical level (Pirapides takes the underspecification of a component as an alternation). Only that information which is different to the one association to the class is actually marked. For instance, in the **Pasiva perifrástica** model associated to the communication class (see Figure 1), the entity element (defined as {entity;**NP**;x;Top:**yes**;no}), has to be realized as an NP and also has to agree with the verb, which is not the usual case in the communication class.

3 Computational Model

LEXPIR allows the construction of patterns for all the possible syntactico-semantic alternations of a verb. However, our goal is to identify the meaning components of these patterns among the components of the partial parsing tree of a sentence. Simultaneous to the selection of the most similar verbal scheme for the sentence, the meaning components are also obtained.

Due to the richness of the language (adjuncts, free order, etc.) there is a need to apply robust pattern recognition techniques which allow to change the position of some elements, the absence of certain elements or the presence of new elements. The following subsections focus on the definition of the technique used for recognizing these complex structures within a sentence.

3.1 Approximate pattern matching

The use of full parsing trees [Atserias et al., 1999] implies previous decisions on the relationship between elements (i.e. PP-attachments). A misidentified syntactic component, or whose limits have not been correctly set, makes difficult not only the recognition of the meaning components but also the recognition of the model itself.

To avoid this problem it was decided to use a syntactic analysis based on syntactic unambiguous groups: chunks [Abney, 1990]. This turns the problem of comparing phrase structure trees into a problem of aligning phrase group sequences.

3.2 Similarity measures

Our similarity measure is defined in terms of the minimum cost sequence of editing operations that transforms one structure into the other. The main differences with previous works on approximate pattern matching based on editing operations [Tsong-li et al., 1994], [Shasha et al., 1994] is that the elements in our sequences are Feature Structures (FS). So the relabelling operation is performed on the attributes.

As a consequence, the following editing operations were defined:

- *Delete*: Deletes an element of the sequence.
- *Insert*: Inserts a new element in the sequence.
- *Move*: Changes the order of an element in the sequence (eg.: "[We] [went] [to Barcelona] [by plane]" and "[by plane] [We] [went] [to Barcelona]").
- *Relabel*: Changes the value of the feature (attribute) of an element in the sequence.

The cost of a sequence of operations is the addition of the cost of each operation. In order to avoid having to choose the smallest model, a correction factor inversely proportional to the number of nodes is added to the similarity measure. It should be pointed out that the number of Relabel and Delete operations gives a measure of the goodness of the matching while the number of Insert operations measures how much information from the sentence is not captured by the pattern.

4 Experiments

The experiments here presented aim to prove not only the feasibility of the linguistic and computational models but also the possibility to apply the system for improving and developing the verbal subcategorization lexicon (LEXPIR).

In order to carry out the experiments a preliminary version of LEXPIR was manually built, which contained 61 verbs belonging to the trajectory class. Then, 170 sentences taken from an Spanish newspaper were labelled by hand with the verbal models and the meaning components. It should be also mentioned that only three sentences present more than one model.

4.1 Processing the corpus

The corpus was pre-processed automatically to obtain a parsed tree for each sentence. Firstly, the corpus was morphologically analized *MACO* [Carmona et al., 1998] and disambiguated *Relax* [Padró, 1997]). Secondly, the Spanish Wordnet [Rodríguez et al., 1998] was used to semantically annotate the corpus with the 79 semantic labels defined in the preliminary version of EuroWordnet Top Ontology. Then, in order to obtain a partial parsing a context free parser based on charts *TACAT* and a wide coverage grammar of Spanish[1] [Castellón et al., 1998] were used to obtain the partial parsing trees (see Figure 1). Finally, those parsed tree were used by our system to produce a case-role representation of the meaning components. For

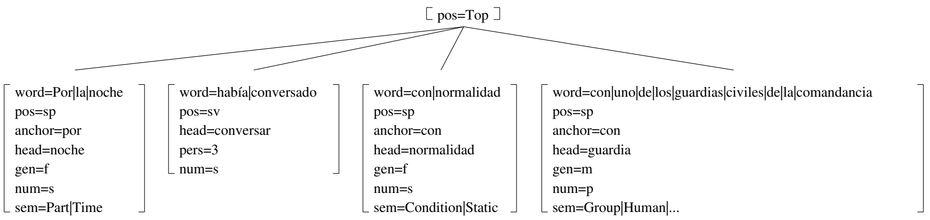

Figure 2: Partial Parse Tree of "At night (He/She) had talked with one of the policemen from the commander's headquarters"

Meaning Comp.	**Lexical Group**
Event	había conversado
Destination	con uno de los guardias civiles de la comandancia

Figure 3: Meaning components obtained with the Basic Model

instance, Figure 3 shows the feature structure of the meaning components obtained from the parsed tree shown in Figure 2.

4.2 Evaluation & Results

The evaluation of a system which performs semantic interpretation is a difficult task. One of the contributions of the MUC's has been to establish a set of evaluation metrics and a common frame for the evaluation of Infor-

[1]Several rules were added to the grammar so as to deal with noun complements.

mation Extraction Systems. The MUC evaluation methodology is based on a pre-alignment of the entities from the solution and response.

However, in order to evaluate the results of our system, the existence of two main differences has to be taken into account: multiple instantiation and entity fragmentation.

- *Multiple instantiation of the same model (entity)*: The generation of different instantiations of the same entity is unusual in Information Extraction, while our system does so. For instance, for the sentence "[Pedro] [habló] [con normalidad] [con Andrés]" (Pedró talked normally with Andrés), two solutions of the basic model are obtained, one filling the role *entity* with *normalidad* and the other with *Andrés*.

- *Entity Fragmentation*: On the other hand, IESs do not always recognize an entity as a whole, so that they generated several entities corresponding to the different fragments. In our system this could not happen as only a model per sentence is considered.

Assuming the existence of only one correct instantiation of a model per sentence, our pre-alignment method consists in comparing all the answers of the same model with the corresponding solution. As in MUC-7, a role is correct if, and only if, both values are equal as strings.

Table 3 shows the results in the identification of the meaning components corresponding to verb arguments and applying the MUC-7 evaluation metrics (see Table 4). Further, Table 2 shows the results obtained on the identification of the verb model . It should be mentioned that due to errors in the pre-processing of the corpus, the system was unable to identify any model for 5 of the 170 sentences.

COR	*INC*	*PRE*	*REC*
158	10	0.94	0.91

Table 2: Model Identification Results

COR	**INC**	**MIS**	**SPU**	**POS**
210	37	89	52	336
ACT	**PRE**	**REC**	**UND**	**OBV**
229	0.6	0.7	0.26	0.17
SUB	**ERR**	**P&R**	**2P&R**	**P&R**
0.15	0.46	0.66	0.64	0.69

Table 3: Meaning Components Results

COR	Number correct	
INC	Number incorrect	
MIS	Number missing	
SPU	Number spurious	
POS	Number possible (elements in the solution)	$COR + INC + MIS$
ACT	Number actual (elements in the response)	$COR + INC + SPU$
REC	Recall	$\frac{COR}{POS}$
PRE	Precision	$\frac{COR}{ACT}$
UND	Undergeneration	$\frac{MIS}{POS}$
OVG	Overgeneration	$\frac{SPU}{ACT}$
SUB	Substitution	$\frac{INC}{COR+INC}$
ERR	Error per response fill	$\frac{INC+SPU+MIS}{COR+INC+SPU+MIS}$
F-MESURES	Weighted combination of REC & PRE	$\frac{(B^2+1.0) \times P \times R}{(B^2 \times P)+R}$

Table 4: MUC-7 Evaluation Metrics

5 Conclusions & Further Work

This paper has presented a semantic parsing approach for non domain-specific texts. Our approach is based on the application of a verbal subcategorization lexicon (LEXPIR) developed in the Pirapides project.

The results of the experiment are very promising. Even though they have been carried out using a limited corpus and lexicon, they have proved the feasibility of the linguistic and computational models.

As further work it is planned to cover linguistic phenomena other than the verbal subcategorization and to expand our system to deal with the combination of multiple models beyond the usual cascade approach. To design a more general framework, it has also been planned to formalize the role identification and model combination processes as a Consistency Labeling

Problem [Padró, 1997] in which different nominal and verbal models can compete for their case-role assignments.

6 Aknowledgments

This research has been partially funded by the Spanish Research Department (Spontaneous-Speech Dialogue System for Limited Domains TIC98-423-C06)

References

Steven Abney. 1990. Parsing by chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-based parsing.*

J. Atserias, I. Castellón, M. Civit, and G. Rigau. 1999. Using diathesis for semantic parsing. In *Venecia per il Trattamento automatico delle lingue (VEXTAL)*, pages 385–392, Venecia, Italy. unipress.

Eric Brill and Raymond J. Mooney. 1997. An Overview of Empirical Natural Language Processing. *Artificial Intelligence Magazine*, 18(14):13–24, Winter. Special Issue on EMpirical Natural Language Processing.

J. Carmona, S. Cervell, L. Márquez, M.A. Martí, L. Padró, R. Placer, H. Rodríeguez, M. Taulé, and J. Turmo. 1998. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Irene Castellón, Montse Civit, and Jordi Atserias. 1998. Syntactic parsing of spanish unrestricted text. In *Proceedings of the 1th Conference on Language Resources and Evaluation (LREC'98)*, Granada. Spain.

A. Fernández and M. A. Martí. 1996. Classification of psycological verbs. *SEPLN*, (20).

A. Fernández, M. A. Martí, G. Vázquez, and I. Castellón. 1999. Establising semantic oppositions for typification of predicates. *Language Design*, (2).

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 1998. University of sheffield: Description of the lasie-ii system as used for muc-7. In *MUC-7.*

Adam Kilgariff. 1997. Foreground and background lexicons and word sense disambiguation for information extraction. In *Proceedings of the Workshop on Lexicon Driven Information Extraction*, Frascati, Italy.

R. Morante, Irene Castellón, and Gloria Vázquez. 1998. Los verbos de trayectoria. In *Proceeding of the conference of the SEPLN.*

Lluis Padró. 1997. *A Hybrid Enviroment for Syntax-semantic Tagging.* Ph.D. thesis, Universitat Politécnica de Catalunya, 15 December.

Ellen Riloff, 1999. *Information Extraction as a Stepping Stone toward Story Understanding.* MIT press, Montreal, Canada.

Horacio Rodríguez, Salvador Climent, Peek Vossen, L. Blocsma, Wim Peters, A. Alonge, F. Bertagna, and A. Rovertini. 1998. The top-down strategy for building euwn: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32(2–3).

D. Shasha, J. Tson-Li Wang, K. Zhang, and Y. Shih. 1994. Exact and approximate algorithms for unorder tree matching. *IEEE transactions on System Man and Cybernetics*, 28(5):668–678, April.

J. Tsong-li, K. Zhang, K. Jeong, and D. Shasha. 1994. A system for approximate tree matching. *IEEE transactions on Knowledge and Data Engineering*, 6(4).

Yorcik Wilks and Roberta Catizone, 1999. *Can We Make Information Extraction More Adaptative*, pages 1–16. Lecture Notes in artificial Intelligence. Springer-Verlang. Subseries of Lectures Notes in Computer Science.

Jordi Atserias is a doctoral student at the Software Department, Universitat Politecnica de Catalunya. *batalla@lsi.upc.es*; URL: *http://www.lsi.upc.es/~batalla*

Irene Castellón is a researcher and professor at the Universitat de Barcelona. *castel@lingua.fil.ub.es*; URL: *http://www.ub.es/ling/labcat.htm*

Montse Civit is a doctoral student at the Software Department, Universitat de Barcelona. *civit@lsi.upc.es*; URL: *http://www.lsi.upc.es/~civit*

*German Rigau*is a researcher and professor at the Software Department, Universitat Politecnica de Catalunya. *g.rigau@lsi.upc.es*; URL: *http://www.lsi.upc.es/~rigau*